arXiv:2508.05115v1 [cs.GR] 7 Aug 2025

RAP: Real-time Audio-driven Portrait Animation with Video Diffusion Transformer

Fangyu Du[1,2*], **Taiqing Li**[1,3*], **Ziwei Zhang**[1*], **Qian Qiao**[1], **Tan Yu**[1], **Dingcheng Zhen**[1], **Xu Jia**[3], **Yang Yang**[2], **Shunshun Yin**[1†], **Siyuan Liu**[1†]

[1]Soul AI, [2]Xi'an Jiaotong University, [3]Dalian University of Technology

{dufangyu, litaiqing, zhangziwei, qiaoqian, yutan, liusiyuan, dingchengzhen, yinshunshun}@soulapp.cn

Project: https://markson14.github.io/RAP

Abstract

Audio-driven portrait animation aims to synthesize realistic and natural talking head videos from an input audio signal and a single reference image. While existing methods achieve high-quality results by leveraging high-dimensional intermediate representations and explicitly modeling motion dynamics, their computational complexity renders them unsuitable for real-time deployment. Real-time inference imposes stringent latency and memory constraints, often necessitating the use of highly compressed latent representations. However, operating in such compact spaces hinders the preservation of fine-grained spatiotemporal details, thereby complicating audio-visual synchronization and increasing susceptibility to temporal error accumulation over long sequences. To reconcile this trade-off, we propose **RAP** (**R**eal-time **A**udio-driven **P**ortrait animation), a unified framework for generating high-quality talking portraits under real-time constraints. Specifically, RAP introduces a hybrid attention mechanism for fine-grained audio control, and a static-dynamic training-inference paradigm that avoids explicit motion supervision. Through these techniques, RAP achieves precise audio-driven control, mitigates long-term temporal drift, and maintains high visual fidelity. Extensive experiments demonstrate that RAP achieves state-of-the-art performance while operating under real-time constraints.

Introduction

Recent advances in latent diffusion models have significantly improved the controllability and quality of portrait animation. Among various control modalities, audio offers continuous and fine-grained temporal cues that naturally align with facial dynamics, making it particularly suitable for driving talking portrait generation. Leveraging these properties, recent audio-driven methods (Xu et al. 2024; Chen et al. 2025) have achieved impressive progress in synthesizing realistic facial expressions and accurate lip movements, greatly enhancing the expressiveness of animated portraits. However, as portrait animation becomes increasingly integrated into interactive scenarios such as virtual communication, digital avatars, and live-streaming, achieving low-latency, high-fidelity generation is critical for delivering smooth and responsive user experiences.

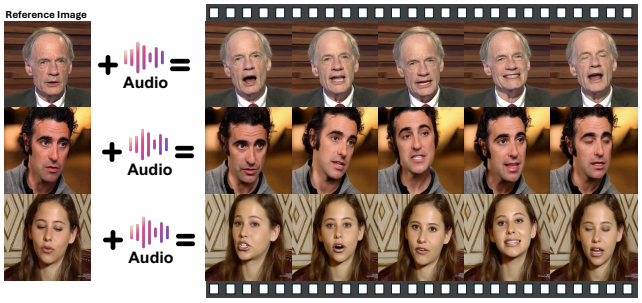


Figure 1: Illustration of the proposed portrait animation framework RAP. Given a reference image and an audio clip, the model generates a natural and vivid talking portrait.

To further improve generation quality and temporal consistency, recent works such as the Hallo (Xu et al. 2024; Cui et al. 2024, 2025) and EchoMimic (Chen et al. 2025; Meng et al. 2025) series introduce sophisticated design strategies. The Hallo series employs dynamic masking and fixed noise injection to guide the model's attention toward facial regions, effectively decoupling local details from global semantics. The EchoMimic series adopts a multi-stage training paradigm that first captures coarse-grained motion and then refines high-frequency details, while incorporating pose control mechanisms to reduce motion drift.

Although effective, existing methods often rely on high-dimensional representations or fine-grained visual storage to maintain coherence and identity consistency. However, this leads to high computational and memory costs, limiting their applicability in real-time settings. Real-time, high-quality audio-driven portrait animation remains challenging due to two main issues: (1) Fine-grained control under high compression. Techniques like LTX-VAE (HaCohen et al. 2024) speed up inference by reducing token length, but increase information density per token, imposing stronger requirements on the diffusion model's ability for fine-grained control. (2) Error accumulation in long sequences. Small pre-

*These authors contributed equally. This work was done while interning at Soul AI.

†Corresponding author, Siyuan Liu is the Project Leader.

diction errors gradually build up over time, causing motion discontinuities, identity drift, and image distortion in longer sequences.

To address the above challenges, we propose RAP, a unified real-time portrait animation framework. RAP aims to simultaneously achieve real-time, high-quality inference performance while effectively mitigating error accumulation issues in long-term generation. Our method is built upon high compression ratio spatiotemporal latent representations to meet strict real-time requirements. To mitigate the difficulty of fine-grained control imposed by highly compressed latent spaces, we innovatively design a hybrid attention mechanism. This mechanism cleverly combines attention to global video coherence with precise control of audio features on fine-grained temporal dimensions over key local video regions (e.g., mouth, eyes). This significantly improves the quality of generating local details under high compression latent space, particularly enhancing the precision of audio-video synchronization.

Furthermore, to effectively solve error accumulation and identity drift issues in long-term generation, RAP proposes a training and inference strategy without explicit motion frame storage. Its core is the innovative static-dynamic hybrid training paradigm. Under this paradigm, we distinguish between static latents and dynamic latents in the VAE latent space. To ensure consistent training and inference, the model is trained to initiate generation both from static latents (for the first clip) and from dynamic latents (for subsequent clips). Rather than applying hard conditioning on preceding outputs, which may lead to the accumulation of errors over time, we adopt a soft guidance mechanism that reuses the denoising process of prior latent features. Through this method, RAP can achieve nearly infinite-length real-time generation while maintaining ID and detail features.

In summary, the main contributions of this paper can be summarized as follows:

- We propose RAP, a novel audio-driven real-time portrait animation generation framework that can generate high-quality, realistic portrait animations while ensuring efficient inference.
- To meet the precision demands of highly compressed latent spaces, we propose a hybrid attention mechanism. It effectively fuses global video context with fine-grained audio cues, enhancing local detail generation and audio-video synchronization.
- To address error accumulation and identity drift in long video generation, we propose a static-dynamic hybrid paradigm with soft latent guidance to support seamless, long-term video generation without explicit motion conditioning.
- We conduct extensive experiments and in-depth analyses to comprehensively validate the effectiveness of the proposed RAP framework. Results demonstrate that RAP achieves high-quality portrait animation and strong temporal coherence under real-time constraints.
- To promote further research in the real-time portrait animation generation field, we will open-source our data cleaning and processing pipeline, as well as complete model training and inference code.

Related Work

Diffusion Models for Video Generation. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have shown strong potential in video generation by modeling complex spatial-temporal dynamics. Early approaches mainly use UNet-based architectures (Blattmann et al. 2023; Bar-Tal et al. 2024; Singer et al. 2022; Wang et al. 2023b, 2024; Wu et al. 2023; Chai et al. 2023; Ceylan, Huang, and Mitra 2023; Guo et al. 2023), where temporal modules are added on top of spatial UNets. However, this separation limits interaction between spatial and temporal features. As a result, these models tend to produce temporal artifacts like frame-wise inconsistencies or flickering pixels, especially in dynamic or textured regions of the video.

To address these limitations, recent advances have shifted toward DiT-based architectures (Yang et al. 2024; Kong et al. 2024; Wan et al. 2025), which treat video frames as spatio-temporal tokens and model them using global self-attention across both dimensions. This unified token-based formulation enables tighter integration of spatial and temporal information, leading to more coherent motion and better generalization to complex video generation tasks. DiT frameworks naturally support cross-modal conditioning—such as text, audio, or pose—by embedding additional control tokens into the same attention space, resulting in superior controllability and alignment. These architectural improvements have contributed to substantial gains in generation quality, positioning DiT-based diffusion models as the current state-of-the-art for high-quality video synthesis.

Audio-driven Portrait Animation. Portrait animation generation aims to produce realistic and temporally coherent facial motions that preserve identity and expression. Audio has become a widely used driving signal due to its rich temporal cues and strong correlation with speech-related dynamics. Early methods (Cheng et al. 2022; Wang et al. 2023a; Gan et al. 2023) often rely on GANs or explicit motion models. For example, Wav2Lip (Prajwal et al. 2020) ensures accurate lip sync via a dedicated discriminator, while SadTalker (Zhang et al. 2023) maps audio to 3DMM parameters for rendering. However, these methods often struggle with expressiveness and temporal consistency, especially in longer sequences.

Recent diffusion-based approaches significantly improve generation quality and coherence. UNet-based models (Xu et al. 2024; Cui et al. 2024; Tian et al. 2024; Wei, Yang, and Wang 2024; Chen et al. 2025) learn audio-to-video mappings in a data-driven manner, enabling more expressive facial motion. For long-form generation, MEMO (Zheng et al. 2024a) and LOOPY (Jiang et al. 2025) model inter-clip dependencies to maintain motion continuity. DiT-based methods (Cui et al. 2025; Wang et al. 2025; Lin et al. 2025) further enhance spatial-temporal coherence and audio-visual alignment through global attention and token-level control. However, their high computational cost remains a challenge for real-time use.

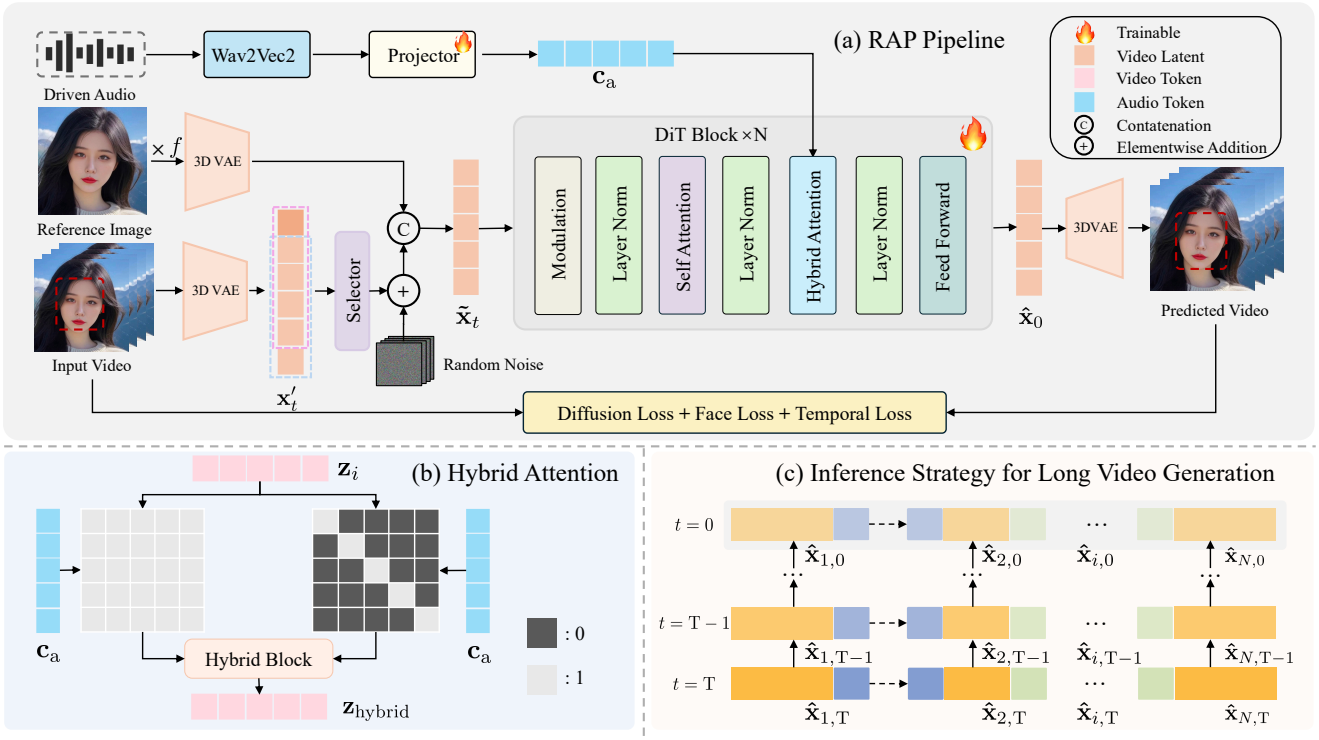


Figure 2: Overview of the proposed RAP framework. (a) Overview of the RAP pipeline, where audio and image inputs are encoded into compressed tokens, followed by DiT-based denoising to generate high-quality portrait videos. (b) The hybrid attention block conducts cross-attention at both short-term and long-term temporal scales, and fuses the results to capture multi-scale audio-motion dependencies. (c) A step-wise inference strategy that progressively guides video generation in the latent space by inheriting information across timesteps.

To address these challenges, we propose RAP (Real-time Audio-driven Portrait Animation). It is a high-performance framework tailored for real-time generation under high compression. RAP employs a hybrid attention mechanism and a static-dynamic training-inference strategy that improve spatiotemporal consistency while keeping latency low.

Methodology

This section systematically presents the methods used in our study. It is organized into four parts: the preliminaries, the RAP framework, a Hybrid Attention mechanism to enhance audio-visual alignment, and a unified training and inference strategy to ensure stable and coherent results.

Preliminary

Diffusion Transformer (DiT). DiT (Peebles and Xie 2023) is a generative model that combines Transformer architectures with diffusion processes, aiming to overcome the structural limitations of traditional U-Net-based latent diffusion models (LDMs) (Rombach et al. 2022). By replacing the U-Net (Ronneberger, Fischer, and Brox 2015) with a Transformer, DiT enhances modeling capacity and scalability across both spatial and temporal dimensions.

In video generation, DiT is often combined with a causal 3D VAE (Kingma and Welling 2013; Zheng et al. 2024b) for spatio-temporal compression. Conditional inputs (e.g., text) are incorporated via adaptive normalization or cross-attention to enable controllable generation. The core objective of DiT is to learn a vector field by minimizing the mean squared error (MSE) between the predicted and ground-truth velocity fields. The training objective is:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{\mathbf{t},\, p_{\mathbf{t}}(\mathbf{x})} \left\| \mathbf{v}_t - \mathbf{u}_t \right\|^2, \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters, $\mathbf{x}_t \sim p_t(\mathbf{x})$ is a sample at time step t, $\mathbf{v}_t$ is the predicted velocity field, and $\mathbf{u}_t = \frac{d\mathbf{x}_t}{dt}$ is the ground-truth velocity. Here, $\mathbf{x}_t$ is obtained by linearly interpolating between a real sample $\mathbf{x}_0$ and Gaussian noise $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, as

$$\mathbf{x}_t = t \cdot \mathbf{x}_1 + (1 - t) \cdot \mathbf{x}_0. \tag{2}$$

We adopt the Wan2.1 Text-to-Video (T2V) model (Wan et al. 2025) with 1.3 billion parameters as our baseline, offering a strong trade-off between performance and computational efficiency in our video generation pipeline.

LTX-VAE. LTX-VAE (HaCohen et al. 2024) is a 3D video VAE that maintains high performance at high compression ratios. To reduce the quadratic computation cost of attention, LTX-VAE downsamples the RGB input with a factor of $(32, 32, 8)$ in space and time, achieving a pixel-to-token

ratio of $1:8192$. This is a $32\times$ higher compression than the commonly used $8 \times 8 \times 4$ scheme, significantly improving efficiency. To compensate for detail loss from heavy compression, LTX-VAE incorporates the final step of a diffusion model in the decoder, enhancing visual quality. To enable real-time video generation, we adopt LTX-VAE in our system. However, although LTX-VAE maintains a comparable static frame performance while drastically reducing time, it still lacks in dynamic frames and temporal continuity, and the increase in the number of frames mapped by its single token increases the difficulty of aligning the details at the frame level in the subsequent network.

RAP

We propose a Real-time Audio-driven Portrait animation model called RAP, which uses a reference image $\mathbf{I}$ and an audio clip $\mathbf{A}$ to generate an identity-consistent portrait animation $\hat{\mathbf{V}}$, as illustrated in Figure 2.

Let $\mathbf{x}_t$ denote the noisy video latent at timestep t. To incorporate identity information, the reference image $\mathbf{I}$ is temporally repeated and then encoded by a variational autoencoder $\mathcal{E}$, yielding a latent representation $\mathbf{x}_{\text{ref}}$ that is channel-aligned with $\mathbf{x}_t$. The two latents are then concatenated along the channel dimension to form a fused representation $\tilde{\mathbf{x}}_t = \text{Concat}(\mathbf{x}_t, \mathbf{x}_{\text{ref}})$.

The audio clip A is encoded by a pretrained Wav2Vec2 model (Baevski et al. 2020), and then passed through a multi-layer perceptron (MLP) to extract temporally aligned audio features $\mathbf{c}_{\text{a}} = \text{MLP}(\text{Wav2Vec2}(\mathbf{A}))$.

Finally, the RAP model $\mathcal{M}$ takes $\tilde{\mathbf{x}}_t$, t, and $\mathbf{c}_{\text{a}}$ as inputs, and predicts the velocity field $\mathbf{v}_t$, which guides the denoising trajectory and ultimately leads to the generation of the audio-driven portrait video. To optimize the model, we propose a composite Flow Matching loss comprising three terms:

$$\mathcal{L} = \mathbb{E}_{\mathbf{t},\, p_{\mathbf{t}}(\mathbf{x})}\Big[\left\|\mathbf{v}_t - \mathbf{u}_t\right\|^2 + \lambda \left\|\mathbf{m} \odot (\mathbf{v}_t - \mathbf{u}_t)\right\|^2 + \mu \left\|\Delta\mathbf{v}_t - \Delta\mathbf{u}_t\right\|^2 \Big], \tag{3}$$

where $\mathbf{v}_t$ and $\mathbf{u}_t$ denote the predicted and ground-truth velocity fields, $\mathbf{m}$ is a facial region mask, and $\odot$ represents element-wise multiplication. The first term (Diffusion Loss) enforces overall motion accuracy, the second term (Face Loss) emphasizes facial motion fidelity, and the third term (Temporal Loss) enforces temporal consistency by minimizing velocity differences across adjacent frames, with $\Delta\mathbf{v_t} = \mathbf{v_t}[:, 1:] - \mathbf{v_t}[:, :-1]$, $\Delta\mathbf{u_t} = \mathbf{u_t}[:, 1:] - \mathbf{u_t}[:, :-1]$. The weights λ and μ control the contribution of the face-focused and temporal regularization terms.

Hybrid Attention

Audio, as a temporally aligned conditional signal, carries rich semantic information. It explicitly governs lip movements at the frame level and implicitly affects global facial expressions and motion intensity. Audio influences these aspects at different temporal scales: lip synchronization requires frame-level alignment, while expressions and motions change more gradually. In high-compression VAEs, each latent spans multiple frames, challenging fine-grained lip alignment alongside overall facial dynamics. To address this, we propose a hybrid attention module that fuses audio and visual latent features at both full-sequence and local-region scales, enabling precise and comprehensive control during generation.

Specifically, we denote the input video tokens to the i-th DiT block as $\mathbf{z}_i \in \mathbb{R}^{(F\times H\times W)\times D}$, where F is the number of latent frames, each frame is divided into $H \times W$ spatial patches, and D denotes the feature dimension used in the Transformer. For simplicity, we omit the time step t in $\mathbf{z}_i$. We patchify the fused latent $\tilde{\mathbf{x}}_t$ to obtain the initial video tokens $\mathbf{z}_0 \in \mathbb{R}^{(F\times H\times W)\times D}$. Meanwhile, the audio features $\mathbf{c}_{\text{a}} \in \mathbb{R}^{(F\times r_f\times N)\times D}$ are extracted as described above, where r_f is the temporal compression ratio of the VAE and N is the number of audio feature layers. These two sequences serve as inputs to the i-th DiT block, where we design two cross-modal fusion mechanisms based on cross-attention (Vaswani et al. 2017).

Full-Sequence Fusion. The global fusion output $\mathbf{z}_{\text{full}}$ is obtained by applying a global cross-attention between the entire sequence of $\mathbf{z}_i$ and $\mathbf{c}_{\text{a}}$:

$$\mathbf{z}_{\text{full}} = \mathbf{z}_i + \text{CrossAttn}(\mathbf{z}_i, \mathbf{c}_{\text{a}}), \tag{4}$$

which enables each video token to fully capture the overall audio-driven emotional and contextual cues, thereby improving temporal coherence in portrait animation.

Fine-grained Window Fusion. The fine-grained fusion output $\mathbf{z}_{\text{window}}$ is computed by performing cross-attention within each latent frame $j \in \{1, \ldots, F\}$, where each spatial video token $\mathbf{z}_i^j \in \mathbb{R}^{(H\times W)\times D}$ (tokens from the j-th latent) attends to each corresponding audio token $\mathbf{c}_{\text{a}}^j \in \mathbb{R}^{(r_f\times N)\times D}$. The results are then concatenated along the frame dimension:

$$\mathbf{z}_{\text{window}} = \mathbf{z}_i + \text{Concat}\left(\text{CrossAttn}(\mathbf{z}_i^j, \mathbf{c}_{\text{a}}^j)\right), \tag{5}$$

which accurately models the correspondence between lip shapes and local articulation in audio, thus improving the alignment between speech and lip motion.

Hybrid Fusion Strategy. Finally, we combine the two fusion branches via a weighted interpolation:

$$\mathbf{z}_{\text{hybrid}} = \alpha(i) \cdot \mathbf{z}_{\text{window}} + (1 - \alpha(i)) \cdot \mathbf{z}_{\text{full}} \tag{6}$$

where the interpolation factor $\alpha(i)$ is defined as:

$$\alpha(i) = \frac{w \cdot i}{L} + \delta, \tag{7}$$

where i denotes the layer index, L denotes the total number of transformer layers, w denotes a scaling parameter, and δ denotes a bias term.

The hybrid attention mechanism enhances lip synchronization and temporal alignment, while preserving semantic consistency throughout the entire video.

Algorithm 1: Training Algorithm for RAP

Require: Encoded video $\mathcal{E}(\mathbf{V})$, Encoded image $\mathbf{x}_{\text{ref}}$, Audio feature $\mathbf{c}_{\text{a}}$

while not converged **do**
  Sample timestep $t \sim \text{Uniform}(\{1, \ldots, T\})$
  Add noise to $\mathcal{E}(\mathbf{V})$ to get $\mathbf{x}'_t$ at timestep t
  Sample $y \sim \text{Bernoulli}(p = \beta)$
  $\mathbf{x}_t = y \cdot \mathbf{x}'_t[:, :k, :, :] + (1 - y) \cdot \mathbf{x}'_t[:, -k:, :, :]$
  $\tilde{\mathbf{x}}_t = \text{Concat}(\mathbf{x}_t, \mathbf{x}_{\text{ref}})$
  Predict the velocity field $\mathbf{v}_t = \mathcal{M}(\tilde{\mathbf{x}}_t, t, \mathbf{c}_{\text{a}})$
  Compute flow matching loss $\mathcal{L}$ according to Eq. 3
  Update θ by minimizing $\mathcal{L}$ using gradient descent
end while

Algorithm 2: Inference Algorithm with Flow Matching

Require: Encoded image $\mathbf{x}_{\text{ref}}$, Audio feature $\mathbf{c}_{\text{a}}$, Generative clip number N, VAE decoder $\mathcal{D}$

for $i = 1$ to N **do**
  Initialize Gaussian noisy latent: $\hat{\mathbf{x}}_{i,\text{T}} \sim \mathcal{N}(0, \mathbf{I_d})$
  for $t = \text{T}$ to 1 **do**
    if $i \neq 1$ **then**
      $\hat{\mathbf{x}}_{i,t} = \text{Concat}(\hat{\mathbf{x}}_{i-1,t}[:, -n:], \hat{\mathbf{x}}_{i,t}[:, n:])$
    end if
    $\tilde{\mathbf{x}}_{i,t} = \text{Concat}(\hat{\mathbf{x}}_{i,t}, \mathbf{x}_{\text{ref}})$
    $\hat{\mathbf{x}}_{i,t-1} = \hat{\mathbf{x}}_{i,t} - \Delta t \cdot \mathcal{M}(\tilde{\mathbf{x}}_{i,t}, t, \mathbf{c}_{\text{a}})$
  end for
  if $i = 1$ **then**
    $\hat{\mathbf{V}}_i = \mathcal{D}(\hat{\mathbf{x}}_{i,0})$
  else
    $\hat{\mathbf{V}}_i = \mathcal{D}(\hat{\mathbf{x}}_{i,0})[:, r_f \cdot (n-1) + 1:]$
  end if
end for

Motion-Frame-Free Training and Inference Strategy for Long Video Generation

In long-form video generation, prior methods often adopt a motion frame strategy, where the last couple frames of the previous clip are used to guide the next. However, this leads to a distribution mismatch: ground-truth motion frames are used during training, while the model relies on its own generated outputs during inference. As generation proceeds, this mismatch accumulates, degrading temporal consistency and visual quality.

To address this issue, we propose a latent inheritance strategy that uses last n intermediate noisy latents from the previous denoising process to softly guide the next clip. Unlike hard guidance based on denoised final results, this approach avoids direct error injection and reduces the risk of accumulated artifacts. By propagating context through latent features rather than fixed outputs, our method enables more stable and coherent long-form video generation.

Nevertheless, applying this strategy to 3D VAE architectures introduces a new challenge. These models typically encode identity from a static initial frame and motion from subsequent dynamic frames. Inherited latents from the previous clip, however, are dynamic and inserted at the start of the next clip—disrupting the original static-dynamic structure and compromising the VAE's encoding.

To resolve this, we propose a dynamic start training scheme. During training, we randomly sample latent features $\mathbf{x}_t$ from the original f-frame $\mathbf{x}'_t$, following a probabilistic strategy: with probability β, from the first a frames (containing both static and dynamic latents), and with probability $1-\beta$, from the last k frames (purely dynamic latents). As a result, this design encourages the model to handle non-static starting conditions. This adaptation ensures compatibility with inherited latents and improves stability across clips. The detailed training and inference procedure is illustrated in Algorithm 1 and Algorithm 2.

Experiments

Datasets and Evaluation Metrics

We constructed the training set from AVSpeech (Ephrat et al. 2018), HDTF (Zhang et al. 2021), VFHQ (Xie et al. 2022), and our own collected data. For preprocessing, we first applied face detection to crop each video frame and discarded samples with resolution below 480×480. All remaining frames were resized to 512×512. We also used a lip-sync consistency metric to remove samples with poor alignment between lip motion and speech. In parallel, we extracted clean human speech from audio using an audio separation tool. After this pipeline, we obtained 222.6 hours of high-quality paired video and audio data. For evaluation, we sampled 75 videos per dataset from HDTF and VFHQ. We compared our method with recent state-of-the-art approaches, including SadTalker (Zhang et al. 2023), Aniportrait (Wei, Yang, and Wang 2024), EchoMimic (Chen et al. 2025), Ditto (Li et al. 2024), and Hallo3 (Cui et al. 2025).

We employ multiple metrics to comprehensively evaluate our method. For visual quality, we adopt Fréchet Inception Distance (FID) (Heusel et al. 2017) to measure the distributional discrepancy between generated and real video frames, and Fréchet Video Distance (FVD) (Unterthiner et al. 2019) to further capture temporal consistency in video generation. To assess the accuracy and smoothness of audio-visual synchronization, we incorporate Sync-C (Chung and Zisserman 2016), which reflects how well the lip motion aligns with the speech content, and Sync-D, which captures the temporal stability of lip dynamics throughout the video. Inference efficiency is reported in FPS (frames per second). All evaluations are conducted on a single NVIDIA A800 GPU.

Implement Details

The model was trained on 32 NVIDIA A800 GPUs using the Adam optimizer. The input video consists of 121 frames with a spatial resolution of 512×512. During training, we randomly select either the first 81 frames (static + dynamic) or the last 88 frames (dynamic only) with a 1:1 probability. Meanwhile, we set a 10% audio dropout to fit the inference classifier-free-guidance (CFG). In training, the batch size of each GPU is 4 and the learning rate is 1×10^{-5}. During inference, we set the CFG scale to 5 to preserve the effectiveness of audio-driven control, and use a latent overlap $n = 3$ to

Figure 3: Qualitative comparison with existing approaches on HDTF and VFHQ dataset. **Videos are available in the Supplementary Material.**

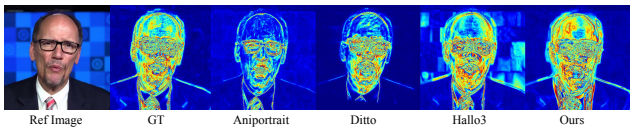


Figure 4: Comparison of temporal consistency and visual drift. Warmer colors denote larger motion amplitude. Our method exhibits minimal background flicker and shift, while preserving significant facial motion.

maintain generation continuity. Our method requires only 8 GB of GPU memory during inference.

Comparison with State-of-the-Art

Quantitative Evaluation. The quantitative results are reported in Table 1 and Table 2. Our method achieves state-of-the-art performance on FVD, Sync-C, and Sync-D, demonstrating strong temporal coherence and superior audio-visual synchronization. While the FID score is slightly inferior to the best-performing baseline—primarily due to the use of highly compressed latent representations that can affect low-level texture fidelity—the gap remains marginal. Additionally, our method runs at real-time inference speed while preserving high perceptual quality.

Qualitative Evaluation. Figure 3 shows visual results generated under identical audio and reference image conditions across different methods. Our method produces lip movements that are highly consistent with the ground truth, reinforcing its advantage in audio-visual alignment. Furthermore, it generates more diverse facial expressions and exhibits a wider range of motion, resulting in more vivid and expressive portrait animations. In contrast, several baseline

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker	21.58	207.67	4.60	9.21	2.17
Aniportrait	19.83	242.29	1.89	11.91	0.69
EchoMimic	**9.00**	155.71	3.56	10.22	0.81
Ditto	12.35	199.13	3.57	10.49	**45.04**
Hallo3	15.95	160.94	3.18	10.72	0.16
Ours	10.24	**122.95**	**4.85**	**8.85**	42.41

Table 1: Quantitative comparison on the HDTF dataset.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
SadTalker	29.80	191.81	4.49	8.78	1.60
Aniportrait	36.58	352.94	1.62	11.73	0.67
EchoMimic	24.69	193.45	2.93	10.30	0.79
Ditto	27.67	254.05	3.31	10.26	**41.24**
Hallo3	23.45	171.00	4.19	9.60	0.11
Ours	**22.68**	**159.93**	**4.78**	**8.40**	39.87

Table 2: Quantitative comparison on the VFHQ dataset.

methods tend to limit motion amplitude to ensure frame stability and temporal smoothness, which often leads to visually static or less engaging results. By contrast, our method maintains a better balance between temporal consistency and motion expressiveness, enabling high-quality generation that remains responsive to audio dynamics. Meanwhile, Figure 4 presents the accumulated inter-frame difference map, where warmer areas correspond to larger motion amplitudes. RAP achieves more expressive facial motions with relatively stable background performance. However, other methods either exhibit significant background flicker or produce nearly static characters with only minor local movements.

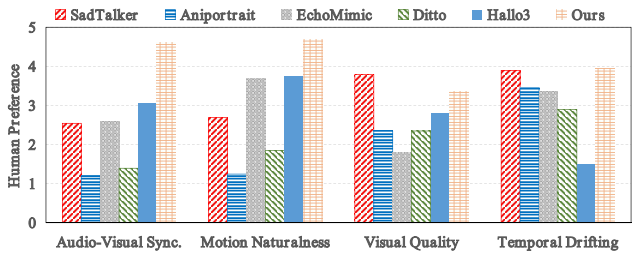


Figure 5: Human preferences among RAP and baselines.

Method	FID↓	FVD↓	Sync-C↑	Sync-D↓
Full-Attention	14.49	196.29	2.13	11.42
Window-Attention	12.35	176.16	**5.03**	**8.68**
Full+Window(2stages)	11.89	166.02	4.81	8.86
Hybrid Attention	**10.24**	**122.95**	4.85	8.85

Table 3: Comparison of different audio injection methods on HDTF dataset.

Human Evaluation. To assess the perceptual quality of the generated videos, we conducted a human study targeting three key dimensions: audio-visual synchronization, naturalness of human body movement, visual quality and resistance to long-term drifting. We selected 40 video samples with varying durations to reflect different temporal challenges, including 25 short-length clips (10–20 seconds), and 15 extended clips (over 2 minutes).

The evaluation involved 127 participants (44.1% aged 18–25, 37.0% aged 25–35,18.9% aged 35–50; 40.2% male, 59.8% female), of whom 73.2% had prior experience with AIGC tools or generative video systems. Participants rated each clip on a 5-point Likert scale in terms of audio-visual synchronization, motion naturalness, visual quality, and robustness to temporal drifting. All videos were presented in randomized order to mitigate ordering bias.

As shown in Figure 5, our approach achieved the highest ratings in terms of audio-visual synchronization, motion naturalness, and robustness to temporal drifting.

Ablation Studies

Hybrid Attention. We conduct a comprehensive ablation study comparing four variants: Full-Attention, Window-Attention, a two-stage approach that applies Full-Attention during initial training followed by Window-Attention for fine-tuning, and our proposed Hybrid Attention mechanism. Among these, only the Hybrid Attention effectively integrates both granularities of control simultaneously. It significantly improves lip-audio alignment while maintaining overall motion consistency and coherence. Additionally, it simplifies training by requiring only a single stage. Detailed results are presented in Table 3.

We further ablate w and δ in Eq. 7, as shown in table 4. We set $w = 1$ and $\delta = 0$ to achieve the best visual quality while preserving high audio-visual synchronization accuracy.

Settings		FID↓	FVD↓	Sync-C↑	Sync-D↓
$w = 0$	$\delta = 0$	14.49	196.29	2.13	11.42
	$\delta = 0.25$	11.31	136.59	3.48	9.94
	$\delta = 0.5$	10.46	126.42	4.80	8.87
	$\delta = 0.75$	10.92	138.97	4.91	8.78
	$\delta = 1$	12.35	176.16	**5.03**	**8.68**
$w = -1$	$\delta = 0.5 - w/2$	10.91	128.69	4.90	8.79
$w = -0.5$		10.79	127.32	4.88	8.80
$w = 0.5$		10.41	124.73	4.82	8.83
$w = 1$		**10.24**	**122.95**	4.85	8.85

Table 4: Comparison of the effects of varying w and δ on hybrid attention.

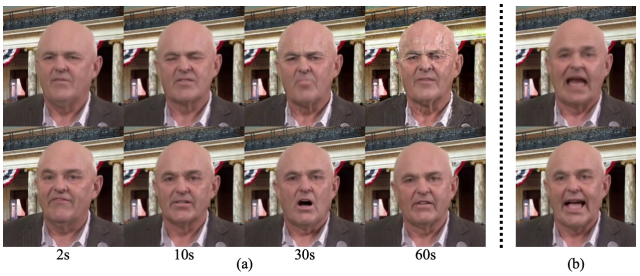


Figure 6: Comparison between inference strategies: subfigure (a) top shows the motion-frame-guided inference, while subfigure (a) bottom shows the inference strategy adopted by RAP. Comparison between training strategies: subfigure (b) top shows training from static latent only, while subfigure (b) bottom shows our hybrid training strategy.

Training and Inference Strategy. Vanilla motion-frame-based approaches directly inject previously generated content as a condition for the subsequent generation process and are trained in a corresponding manner. This leads to a strong dependency on the input motion frames and causes the model to inherit and accumulate errors from them. In contrast, our proposed training and inference strategy utilizes the preceding generated results solely to guide the denoising process of the next generative clip, thereby circumventing the conventional teacher-forcing dilemma. As shown in Figure 6 (a), with extended inference duration, the results from the motion-frame baseline rapidly develop conspicuous artifacts that progressively accumulate. Conversely, RAP's results exhibit no significant degradation as the duration increases. Our tests demonstrate that a one-hour-long video result can maintain the same quality as its initial segments. Furthermore, our mixed static-dynamic training is specifically designed to complement the aforementioned inference scheme. Unlike traditional methods that train exclusively on static-to-dynamic pairs, our strategy also incorporates dynamic-to-dynamic pairs. This approach resolves the inheritance inconsistency issue that typically emerges from the second window onward, thus achieving superior transitional performance, as Figure 6 (b) illustrated.

Conclusion

In this work, we propose RAP, a real-time audio-driven portrait animation framework. By introducing a hybrid attention mechanism and a static-dynamic joint training-inference strategy, RAP achieves precise alignment between audio and visual content under highly compressed representations, enabling the real-time generation of natural and coherent long-term portrait animations.

Under rapid motion scenarios, the use of a high compression ratio VAE may still lead to motion blur and ghosting artifacts due to latent information loss, limiting the fidelity of generated results. Furthermore, extending the framework to real-time multi-speaker conversations and dynamic scene generation remains an important direction for future work. In addition, exploring the applicability of our training and inference strategy to other modality-guided portrait animation tasks, and more broadly to general video generation, is a promising avenue for future research.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; and Ma, C. 2025. EchoMimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2403–2410.

Cheng, K.; Cun, X.; Zhang, Y.; Xia, M.; Yin, F.; Zhu, M.; Wang, X.; Wang, J.; and Wang, N. 2022. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Chung, J. S.; and Zisserman, A. 2016. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, 251–263. Springer.

Cui, J.; Li, H.; Yao, Y.; Zhu, H.; Shang, H.; Cheng, K.; Zhou, H.; Zhu, S.; and Wang, J. 2024. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*.

Cui, J.; Li, H.; Zhan, Y.; Shang, H.; Cheng, K.; Ma, Y.; Mu, S.; Zhou, H.; Wang, J.; and Zhu, S. 2025. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21086–21095.

Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W. T.; and Rubinstein, M. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*.

Gan, Y.; Yang, Z.; Yue, X.; Sun, L.; and Yang, Y. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22634–22645.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. 2024. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jiang, J.; Liang, C.; Yang, J.; Lin, G.; Zhong, T.; and Zheng, Y. 2025. Loopy: Taming Audio-Driven Portrait Avatar with Long-Term Motion Dependency. arXiv:2409.02634.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Li, T.; Zheng, R.; Yang, M.; Chen, J.; and Yang, M. 2024. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. *arXiv preprint arXiv:2411.19509*.

Lin, G.; Jiang, J.; Yang, J.; Zheng, Z.; and Liang, C. 2025. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*.

Meng, R.; Zhang, X.; Li, Y.; and Ma, C. 2025. EchoMimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5489–5498.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. A lip sync expert is all you need for

speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, 484–492.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tian, L.; Wang, Q.; Zhang, B.; and Bo, L. 2024. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, 244–260. Springer.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, J.; Qian, X.; Zhang, M.; Tan, R. T.; and Li, H. 2023a. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14653–14662.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, M.; Wang, Q.; Jiang, F.; Fan, Y.; Zhang, Y.; Qi, Y.; Zhao, K.; and Xu, M. 2025. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*.

Wang, X.; Zhang, S.; Yuan, H.; Qing, Z.; Gong, B.; Zhang, Y.; Shen, Y.; Gao, C.; and Sang, N. 2024. A recipe for scaling up text-to-video generation with text-free videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6572–6582.

Wei, H.; Yang, Z.; and Wang, Z. 2024. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7623–7633.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 657–666.

Xu, M.; Li, H.; Su, Q.; Shang, H.; Zhang, L.; Liu, C.; Wang, J.; Yao, Y.; and Zhu, S. 2024. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; and Wang, F. 2023. SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8652–8661.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3661–3670.

Zheng, L.; Zhang, Y.; Guo, H.; Pan, J.; Tan, Z.; Lu, J.; Tang, C.; An, B.; and Yan, S. 2024a. MEMO: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024b. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.

Appendix

Dataset processing

The training data used in this work includes AVSpeech (Ephrat et al. 2018), HDTF (Zhang et al. 2021), VFHQ (Xie et al. 2022), and our own collected video dataset. Since AVSpeech and our collected data contain a large number of low-quality samples (e.g., low resolution or audio-visual misalignment), we apply a series of preprocessing steps to ensure the quality of audio-driven face generation. Specifically, we first compute the Sync-C and Sync-D metrics for each video and retain only those with Sync-C > 1 and Sync-D < 13. We then perform face detection on the first frame and discard videos in which the detected face region is smaller than 480×480. For the remaining videos, the face bounding box is expanded by a factor of two to ensure that the face occupies approximately half of the cropped frame width, followed by resizing to 512×512 to meet the network input requirements. All audio tracks are converted to mono and undergo offline voice extraction, with the resulting clean speech saved as `.pt` files for efficient training. Moreover, we discard videos shorter than 5 seconds to ensure sufficient sequence length for training with both static and dynamic frames. After preprocessing, the total durations of the datasets are as follows: AVSpeech 121.6 hours, our own dataset 81.39 hours, HDTF 6.73 hours, and VFHQ 12.86 hours.

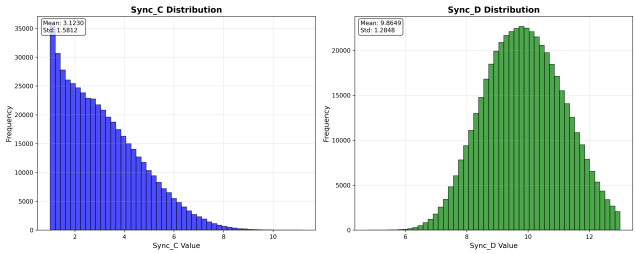


Figure 7: Distribution of Sync-C and Sync-D in raw datasets.

Experiments

Experimental Setup Details. In RAP, we adopt the LTX-VAE (HaCohen et al. 2024) together with the Wan2.1 (Wan et al. 2025) T2V model containing 1.3B parameters. Specifically, the raw video frames $\mathbf{V} \in \mathbb{R}^{3\times121\times512\times512}$ are compressed by LTX-VAE with a spatial-temporal compression factor of $(8, 32, 32)$ to get $\mathbf{x}_0 \in \mathbb{R}^{128\times16\times16\times16}$. We add Gaussian noise to $\mathbf{x}_0$ based on a sampled timestep $t \sim \mathcal{U}(0, T)$, obtaining the noisy latent $\mathbf{x}'_t \in \mathbb{R}^{128\times16\times16\times16}$, from which we randomly sample latent features $\mathbf{x}_t \in \mathbb{R}^{128\times11\times16\times16}$ following a probabilistic strategy: with probability 0.5, from the first 11 frames (containing both static and dynamic latents), and with probability 0.5, from the last 11 frames (purely dynamic latents). To preserve identity, we extract the corresponding latent features $\mathbf{x}_{\text{ref}}$ from the reference image and concatenate them with $\mathbf{x}_t$ along the channel dimension, yielding the fused input $\tilde{\mathbf{x}}_t \in \mathbb{R}^{256\times11\times16\times16}$.

This fused latent $\tilde{\mathbf{x}}_t$ is reshaped and processed by a patchify layer with kernel size $(1, 1, 1)$, resulting in a token sequence $\mathbf{z}_0 \in \mathbb{R}^{(11\times16\times16)\times1536}$ to match the input dimension of the Transformer. We adopt a 30-layer Transformer with 12 attention heads and a feed-forward network (FFN) hidden dimension of 8960 to enhance expressive capacity.

The output sequence is then depatchified to reconstruct the latent $\hat{\mathbf{x}}_0 \in \mathbb{R}^{128\times11\times16\times16}$, and decoded by the VAE decoder to generate video frames $\hat{\mathbf{V}} \in \mathbb{R}^{3\times81\times512\times512}$.

The impact of CFG scale. Classifier-free-guidance (CFG) is a crucial technique for enhancing controllability during model inference. We conduct an ablation study on different CFG scales, and the results are shown in the table 5 below.

CFG scale	FID↓	FVD↓	Sync-C↑	Sync-D↓
2	11.50	123.94	4.61	9.07
4	10.33	**120.48**	<u>4.83</u>	**8.85**
5	<u>10.22</u>	<u>122.95</u>	**4.85**	**8.85**
6	**10.08**	126.55	4.79	<u>8.91</u>
8	10.89	130.31	4.71	8.98

Table 5: Ablation of CFG scale on the HDTF dataset. We set it to 5 to balance visual quality and audio-visual synchronization performance.

overlap length	FID↓	FVD↓	Sync-C↑	Sync-D↓	FPS↑
$n = 1$	10.61	125.63	**4.89**	**8.82**	**52.96**
$n = 2$	10.38	<u>123.58</u>	<u>4.86</u>	8.87	<u>46.71</u>
$n = 3$ (Ours)	<u>10.24</u>	**122.95**	4.85	<u>8.85</u>	42.41
$n = 4$	**10.22**	123.67	4.83	8.89	38.12

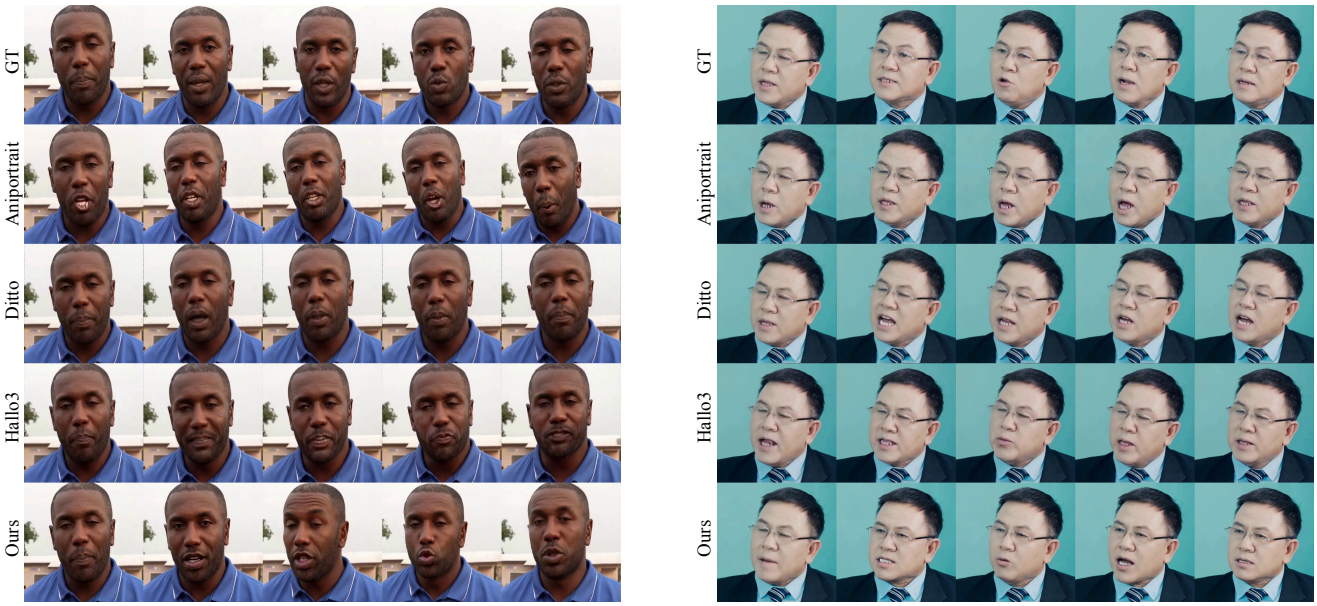
Table 6: Ablation of overlap length on the HDTF dataset.

The impact of overlap length. In our approach, we adopt a temporal guidance strategy, where the last n latent frames of the preceding clip are reused as the first n latents of the next clip to guide generation. We conduct an ablation study on the number of overlapping latent frames n to investigate its impact on generation quality, audio-visual synchronization, and inference speed. In the final setting, we choose $n = 3$ as it achieves the best trade-off between visual fidelity, sync accuracy, and efficiency.

More qualitative results. More qualitative results are illustrated in Figures 8, 9

(a) (b)

Figure 8: Qualitative results on HDTF dataset



(a) (b)

Figure 9: Qualitative results on VFHQ dataset